Journal of Advance Research in Mathematics and Statistics (ISSN: 2208-2409)

DOI: https://doi.org/10.53555/nnms.v3i6.552 **Publication URL: http://nnpub.org/index.php/MS/article/view/552**

DEVELOPING MODELS TO EXPLAIN POINT SPREAD OF NCAA WOMEN'S DIVISION II BASKETBALL GAMES

Feifei Huang[1] and Rhonda Magel[2]*

*[1,2]*Department of Statistics North Dakota State University Fargo, ND 58108*

Corresponding Author: -

Abstract: -

A model is developed using in-game statistics to help explain the final point spread of a Women's Division II Basketball Tournament game. A model is also developed to estimate the probability of a team winning the game given significant in-game statistics. Both of the models are verified based on a random sample of basketball games that were not used in the development of the models. The models were then used to predict the outcomes of the 63 games played in the 2015 NCAA Division II Women's Basketball tournament replacing the actual in-game statistics with seasonal averages of the corresponding statistics. Results are given.

Keywords: *Least Squares Regression, Logistic Regression, Basketball Tournament, In-Game Statistics*

I.

1. INTRODUCTION

There has been much attention paid to the NCAA Division I Men's Tournament held at the end of March and into April of each year, also known as "March Madness". After the first round of "March Madness", there are 64 teams split into 4 regions that are in the tournament. It was estimated that in 2013, more than 100 million people worldwide participated in filling out brackets to make predictions on the games (Barra, 2014). Many researchers have worked on developing models, and/or, ratings to help predict the results of" March Madness" including Dirks (2000), Kubatko, Oliver, Pelton, and Rosenbaum (2007); Magel and Unruh (2013), Pomeroy (2014), Sagarin (2014), Schwertman, McCready, and Howard (1991) Shen, Hua, Zhang, Mu, and Magel (2015); and West (2006,2008), among others.

Wang and Magel (2014) developed least squares regression models and logistic regression models for various rounds of the NCAA Division I Women's Basketball tournament using in-game statistics. The models were validated and then used for predicting the results of the 2014 NCAA Division I Women's Basketball Tournament. When seasonal averages were substituted for the in-game statistics, the least squares regression models predicted 87.5%, 81.3%, and 73.3% of the games correctly for the first, second and third and higher rounds, respectively. The logistic models predicted 90.63%, 81.25%, and 73.33% of the games correctly for the first, second, and higher rounds, respectively.

Not as much interest has been paid in the past to the Division II Women's Basketball tournament. In this study, we would like to develop a least squares regression model that helps identify the in-game statistics that explain the variation in point spread of a NCAA Women's Division II basketball game. We would also like to develop a logistic regression model that helps to estimate the probability that a particular team will win the game based on differences of in-game statistics found to be significant.

There are approximately 300 schools in Division II located in both the United States and Canada (NCAA, "About NCAA Division II", 2015.). Sixty-four women's basketball teams from these schools play in the annual tournament. Twenty-four teams get an automatic entry because of winning their conference's championship. The remaining 40 teams are decided upon by the NCAA selection committee (NCAA, "About NCAA Division II",2015).

The 64 teams are divided into 8 regions: Atlantic, East, Central, Midwest, South Central, South, Southeast, and West (NCAA, "Pre-Championship" 2015-2016). Each region has teams seeded from 1 to 8 with the strongest teams in each region being seeded as 1. The teams with seeds summing up to 9 will play each other in the first round. The teams that win in the first round will advance to the second round, and the process continues.

The winner of each region advances to the Women's Elite Eight. There are a total of 6 rounds played with the 6^{th} round being the Championship game (NCAA, *Championships*).

2. Model Development

In order to develop a model to help explain the variation in point spread of a tournament game and develop a logistic regression model to help estimate the probability of a particular team winning the game, results and in-game statistics were collected from the 2012, 2013, and 2014 Women's Division II Basketball tournaments (NCAA *2012*, NCAA *2013*, NCAA *2014*). For each game played in these tournaments, we collected the ingame statistics for each team that are given in Table 2.1.

Table 2.1: All In-Game Statistics Collected for Building Models

Total points (TP) (dependent variable)	Defensive rebounds (DE)
Field goal percentage (FG%)	Personal fouls (PF)
Three-point field goal percentage (3PT%)	Assists (A)
Free throw percentage (FT%)	Turnovers (TO)
Offensive rebounds (OF)	Blocks (BLK)

We randomly selected one team playing in the game to be the "team of interest". The other team was then the "opposing team". Differences for all the in-game statistics in Table 2.1 were found in the order of the "team of interest" minus the "opposing team". The dependent variable in the model using least squares regression was the difference in total points scored between the two teams. The differences of all the other in-game statistics collected from both teams were the independent variables considered for possible entry into the model. Stepwise selection with a significance level of 0.10 for entry and exit into the model was used initially to help develop this model. The intercept term in the model was set to zero since it was assumed that if all of the in-game statistics for both teams were the same, the point spread should be zero on average. The order in which teams were considered in the model should not matter.

Seven of the possible nine variables considered were found to be significant. Variance Inflation Factors (VIFs) associated with each of these variables were calculated to see if there were problems with multicollinearity (Abraham and Ledolter, 2006). The VIFs, the estimated coefficients associated with the variables, their standard errors, and associated p-values are given in Table 2.2.

Table 2.2: Point Spread Model Coefficient Estimates

Variable	Coefficient Estimate	Standard Error	T-Value	P-Value	VIF
diff_FG%	0.9949	0.0432	23.01	0.000	2.87
diff_3PT%	0.1498	0.0209	7.16	0.000	1.43
diff_FT%	0.2082	0.0180	11.54	0.000	1.20
diff_OF	0.8387	0.0588	14.26	0.000	1.54
diff_PF	-0.4525	0.0662	-6.84	0.000	1.22
diff_A	0.1775	0.0645	2.75	0.007	1.87
diff_TO	-0.7881	0.0521	-15.11	0.000	1.16

All of the VIFs were less than 10 indicating that multicollinearity should not be a problem and we should be able to interpret the coefficients (Abraham and Ledolter, 2006). The variables remaining in the model were highly significant with the largest p-value being 0.007. The R^2 for the model was 92.49%, with the adjusted R^2 equal to 92.19%. The predictive R^2 was found to be 91.66% indicating that the model should be able to estimate the point spread of a game well when the in-game statistics are known.

The point spread model is given below:

$$\widehat{diff\_PT} = 0.9949\, diff\_FG\% + 0.1498\, diff\_3PT\% + 0.2082\, diff\_FT\% + 0.8387\, diff\_OF - 0.4525\, diff\_PF + 0.1775\, diff\_A - 0.7881\, diff\_TO$$

The following statistics have positive coefficients associated with them which is to be expected: diff_FG%, diff_3PT%, diff_FT%, diff_OF, diff_A. It is noted that each additional field goal percentage over the opposing team is estimated to be worth on average approximately 1 point. Each additional rebound over the other team is worth approximately 0.84 points. The following statistics have negative coefficients associated with them: diff_PF and diff_TO. Each additional turnover compared to the opposing team, costs the team on average 0.79 points over the opposing team, and each additional personal foul compared to the opposing team, costs the team an average of 0.45 points.

Residual plots were found for the point spread model to check the assumptions of the errors being independent, approximately normally distributed, with a mean of zero, and independent and are given in Figure 4.1. The assumptions appear to be satisfied.

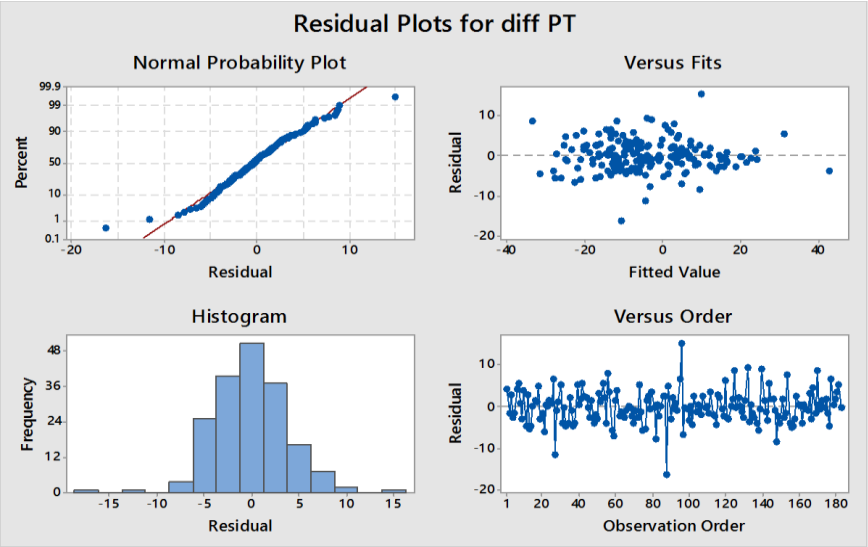

Figure 4.1: Residual Plots for Point Spread Model

Stepwise selection was also initially used in relation to the logistic model with and alpha level of 0.10 of entry and exit into the model. In this case, six of the nine variables considered were found to be significant in the model. The parameter estimates and tests for each of these variables is given in Table 2.3.

Table 2.3: Logistic Regression Model Coefficient Estimates

Parameter	DF	Estimate	Standard Error	WaldChiSquare	P_Value
diff_FG%	1	0.4987	0.1038	23.0761	<.0001
diff_3PT%	1	0.0798	0.0335	5.6787	0.0172
diff_FT%	1	0.1082	0.0285	14.4493	0.0001
diff_OF	1	0.4360	0.1035	17.7308	<.0001
diff_PF	1	-0.2538	0.0833	9.2787	0.0023
diff_TO	1	-0.3704	0.0902	16.8547	<.0001

The developed logistic regression model estimating the probability that the "team of interest" will win the game given the differences in in-game statistics was found to be:

$$\pi(x_i) = \frac{e^{0.4987\,diff\_FG\% + 0.0798\,diff\_3PT\% + 0.1082\,diff\_FT\% + 0.4360\,diff\_OF - 0.2538\,diff\_PF - 0.3704\,diff\_TO}}{1 + e^{0.4987\,diff\_FG\% + 0.0798\,diff\_3PT\% + 0.1082\,diff\_FT\% + 0.4360\,diff\_OF - 0.2538\,diff\_PF - 0.3704\,diff\_TO}}$$

3. Model Validation

If we actually knew in-game statistics from basketball games that were not used in the development of the models and our models correctly determined which team had won the game a large percentage of the time, we would consider our models to be validated. The in-game statistics found to be significant in the models were collected from all of the 63 games played in the 2015 tournament in order to validate the models (NCAA *2015*). For each of the 63 games in this tournament, we placed the significant in-game statistic differences into the point spread model. If we found the estimated difference in point spread to be positive, we predicted that the "team of interest" won the game. If we found the estimated difference in point spread to be negative, we predicted that the "team of interest" lost the game. Our predictions were actually compared to what happened in the game. As an example, the in-game statistics were collected in the basketball game between Lewis (IL.) and Limestone (S.C.) played on March 24, 2015 and are given in Table 3.1. The differences were placed in the point spread model.

Table 3.1: Example for Independent Variable Data Entry in point spread model (Lewis (IL.) vs. Limestone (SC.) on 3/24/2015)

Significant Statistical Measures	Game Results for Lewis ("Team of interest")	Game Results for Lime Stone ("Opposing team")	Differences (Significant Variable Values)	Significant Independent Variable Names
FG%	34.4	36.5	34.4-36.5 = -2.1	diff_FG%
3PT%	30	41.7	30-41.7 = -11.7	diff_3PT%
FT%	83.3	83.3	83.3- 83.3 = 0	diff_FT%
OF	16	13	16-13 = 3	diff_OF
PF	14	10	14-10 = 4	diff_PF
A	14	6	14-6 = 8	diff_A
TO	20	15	20-15 = 5	diff_TO

$$\begin{aligned}\widehat{diff\_PT} &= 0.9949\,diff\_FG\% + 0.1498\,diff\_3PT\% + 0.2082\,diff\_FT\% \\ &\quad + 0.8387\,diff\_OF - 0.4525\,diff\_PF + 0.1775\,diff\_A - 0.7881\,diff\_TO \\ &= 0.9949 * (-2.1) + 0.1498 * (-11.7) + 0.2082 * (0) + 0.8387 * 3 - \\ &\quad 0.4525 * 4 + 0.1775 * 8 - 0.7881 * 5 \\ &= -5.65635\end{aligned}$$

Since the predicted point spread is less than zero, this game was coded as a loss for Lewis, who actually lost the game by a score of 58 to 61, or a point difference of -3. This was done for each of the 63 games in the 2015 tournament and the accuracy of the point spread model is given in Table 3.2.

Table 3.2: Accuracy of Point Spread Model Using in-game Statistics

Point spread		Predicted		
Actual		Win	Loss	Total
	Win	23	2	25
	Loss	1	37	38
	Total	24	39	63
	Overall Accuracy			95.24%

The estimated probability of the "team of interest" winning the game was calculated for each of the 63 games by placing the actual in-game statistic differences found to be significant in the logistic model. If the model determined the probability that the "team of interest" would win would be greater than 0.5, a win was predicted for the "team of interest". Otherwise, a loss was predicted for the "team of interest". Consider again, the game between Lewis and Limestone (Table 3.1)

Using the logistic regression model, Lewis had a projected probability of victory of:

$$\pi(x_i) = \frac{e^{0.4987\ diff_{FG}\% + 0.0798\ diff_{3PT}\% + 0.1082\ diff_{FT}\% + 0.4360\ diff_{OF} - 0.2538\ diff_{PF} - 0.3704\ diff_{TO}}}{1 + e^{0.4987\ diff_{FG}\% + 0.0798\ diff_{3PT}\% + 0.1082\ diff_{FT}\% + 0.4360\ diff_{OF} - 0.2538\ diff_{PF} - 0.3704\ diff_{TO}}}$$

$$= \frac{e^{0.4987*(-2.1) + 0.0798*(-11.7) + 0.1082*(0) + 0.4360*3 - 0.2538*4 - 0.3704*5}}{1 + e^{0.4987*(-2.1) + 0.0798*(-11.7) + 0.1082*(0) + 0.4360*3 - 0.2538*4 - 0.3704*5}}$$

$$= 0.0284$$

Since this projected probability of victory is less than 0.50, this game is coded as a predicted loss for Lewis and recall that Lewis did lose the game. This process was then repeated for the sample of 63 games in the 2015 tournament, and the accuracy of the logistic regression model is given in Table 3.3.

Table 3.3: Accuracy of Logistic Regression Model Using in-game Statistics

Point spread		**Predicted**		
Actual		Win	Loss	Total
	Win	23	2	25
	Loss	1	37	38
	Total	24	39	63
	Overall Accuracy			95.24%

Both models had the same estimated accuracy of 95.24%. Since the models did correctly predict the winner in over 90% of the games when the in-game statistics were known, the models were considered validated.

4. Model Prediction

We next tried to use the models to make predictions as to which team would win the game ahead of time without knowing the actual in-game statistics. This was done using all 63 games in the 2015 tournament. In place of the actual in-game statistics, the 2015 seasonal averages were found for each of the significant in-game statistisc and differences in these averages were placed into the model in the order of "team of interest" minus the "opposing team" instead of using the actual in-game statistics differences. As an example, we will consider the game played between Union and West Florida. The seasonal averages for both Union and West Florida for each of the significant in-game statistics are given in Table 4.1 and the differences are taken.

Table 4.1: Example in Model prediction

(Union (TN.) vs. West Florida (FL.) on 3/14/15)

Significant Statistical Measures	**Seasonal Averages for Union ("Team of interest")**	**Seasonal Averages for West Florida ("Opposing team")**	**Differences in Seasonal Averages**	**Significant Independent Variable Names**
FG%	44.9	38.1	44.9-38.1=6.8	diff_FG%
3PT%	38	29.6	38-29.6=8.4	diff_3PT%
FT%	79.3	68.8	79.3-68.8=10.5	diff_FT%
OF	9	17	9-17=-8	diff_OF
PF	15.6	18.7	15.6-18.7=-3.1	diff_PF
A	14.7	10.9	14.7-10.9=3.8	diff_A
TO	11.7	17.8	11.7-17.8=-6.1	diff_TO

Using the least squares regression model already developed, Union had a predicted point spread of:

$$\widehat{diff\_PT} = 0.9949\ diff\_FG\% + 0.1498\ diff\_3PT\% + 0.2082\ diff\_FT + 0.8387\ diff\_OF - 0.4525\ diff\_PF + 0.1775\ diff\_A - 0.7881\ diff\_TO$$

$$= 0.9949*(6.8) + 0.1498*(8.4) + 0.2082*(10.5) + 0.8387*(-8) - 0.4525*(3.1) + 0.1775*(3.8) - 0.7881*(-6.1)$$

$$= 10.3848$$

Since the predicted point spread is greater than zero, this game was coded as a predicted win with a predicted point spread of 10.4 for Union, who actually won the game by a point spread of 9 versus West Florida.
Using the logistic regression model, Union had a projected probability of victory of:

$$\pi(x_i) = \frac{e^{0.4987\ diff_{FG}\%+0.0798\ diff_{3PT}\%\ +0.1082\ diff_{FT}\%\ +0.4360\ diff_{OF}-0.2538\ diff_{PF}-\ 0.3704\ diff_{TO}}}{1+e^{0.4987\ diff_{FG}\%+0.0798\ diff_{3PT}\%\ +0.1082\ diff_{FT}\%\ +0.4360\ diff_{OF}-0.2538\ diff_{PF}-\ 0.3704\ diff_{TO}}}$$

$$= \frac{e^{0.4987*(6.8)+0.0798\ *(8.4)\ +0.1082*(10.5)\ +0.4360*(-8)-0.2538\ *(-3.1)\ -\ 0.3704*(-6.1)}}{1+e^{0.4987*(6.8)+0.0798\ *(8.4)\ +0.1082*(10.5)\ +0.4360*(-8)-0.2538\ *(-3.1)\ -\ 0.3704*(-6.1)}}$$

$= 0.9915$

The projected probability of victory is 0.9915, which is greater than 0.50, so the game is a correctly predicted win for Union.

This process was then repeated for the sample of 63 games in the NCAA Division II 2015 Women's Basketball Tournament (NCAA *2015*), with the number of predicted victories and defeats from both models separately being compared to the actual number of victories and defeats. The accuracy is calculated in Table 4.2 and Table 4.3 for the point spread and the logistic models, respectively. The point spread model correctly predicted 61.90% of the games when the difference in seasonal averages were used in place of the actual ingame statistics. The logistic model correctly predicted 65.08% of the games.

Table 4.2: Prediction Accuracy of Point Spread Model

Point spread			**Predicted**	
Actual		Win	Loss	Total
	Win	10	14	24
	Loss	10	29	39
	Total	20	43	63
	Overall Accuracy			61.90%

Table 4.3: Prediction Accuracy of Logistic Regression Model

Point spread			**Predicted**	
Actual		Win	Loss	Total
	Win	11	13	24
	Loss	9	30	39
	Total	20	43	63
	Overall Accuracy			65.08%

5. Conclusions

Two models were developed for use with NCAA Division II Women's Basketball games. One of these was a point spread model that explained the variation in point spread of a women's basketball game based on knowing the differences of 7 in-game statistics for the two teams. The other model was a logistic regression model that estimated the probability of the "team of interest" winning the game if the differences of 6 in-game statistics were known. Both of the models were validated. If the actual in-game statistics were known, both models had an estimated accuracy of 95.24% of being able to name the correct winner of the game.

The models were then used to try predicting the results of the 63 games in the 2015 Division II Women's Basketball tournament. Seasonal averages of the significant in-game statistics were found for each of the two teams playing against each other for each game. The differences of these seasonal averages were placed in each model in place of the actual in-game statistics and predictions were made as to which team would win the game. The point spread and logistic models had accuracies of 61.90% and 65.08 %, respectively, when seasonal averages were used in place of the in-game statistics. These accuracies are better than flipping a coin, but we would like to improve them. When in-game statistics are known, the models do very well. Possible future research will include examining better ways of estimating the in-game statistics. This could involve estimating the in-game statistics using the second half seasonal averages instead of the seasonal averages for the entire season, or using a three or four game moving average of the statistics found to be significant in the models.

REFERENCES

[1].Abraham, Bovas, and Ledolter, Johannes (2006). *Introduction to Regression Modeling.* Belmont: Thomson Brooks/Cole, 2006. Print.

[2].Barra, Allen (2014). *Is March Madness a Sporting Event—or a Gambling Event?* Retrieved from http://www.theatlantic.com/entertainment/archive/2014/03/is-march-madness-a-sporting- event-or-a-gambling-event/284545 on 12/8/2015.
[3].Dirks, Kurt T. (2000). *Trust in Leadership and Team Performance: Evidence from NCAA Basketball.* Journal of Applied Psychology, 85(6): 1004-1012.
[4].Kubatko, Justin; Oliver,,Dean; Pelton,Kevin; and Rosenbaum, Dan T. (2007). *A Starting Point for Analyzing Basketball Statistics.* Journal of Quantitative Analysis in Sports, 3(3): Article 1.
[5].Magel, Rhonda, and Unruh, Samuel (2013). *Determining Factors Influencing the Outcome of College Basketball Games.* Open Journal of Statistics (http://www.scirp.org/journal/ojs).
[6].NCAA. 2015. *2016 Division II Women's Basketball Championship: Pre-Championship 2015-16 Manual.* Retrieved from http://www.ncaa.org/sites/default/files/201516_DIIWBB_PreChamps_20151109.pdf on 12/8/2015.
[7].NCAA. *2012 DII Women's Basketball Championship*. Retrieved from http://fs.ncaa.org/Docs/stats/w_basketball_champs_records/2012/d2/html/confstat.htm on 12/9/2015.
[8].NCAA. *2013 DII Women's Basketball Championship*. Retrieved from http://fs.ncaa.org/Docs/stats/w_basketball_champs_records/2013/d2/html/confstat.htm on 12/9/2015.
[9].NCAA. *2014 DII Women's Basketball Championship*. Retrieved from http://fs.ncaa.org/Docs/stats/w_basketball_champs_records/2014/d2/html/confstat.htm on 12/9/2015.
[10]. NCAA. *2015 DII Women's Basketball Championship*. Retrieved from http://fs.ncaa.org/Docs/stats/w_basketball_champs_records/2015/d2/html/confstat.htm on 12/9/2015.
[11]. NCAA. *About NCAA Division II*. Retrieved from http://www.ncaa.org/about?division=d2 on 12/8/2015.
[12]. NCAA. *Championships*. Retrieved from http://www.ncaa.org/about/what-we-do/championships on 12/8/15.
[13]. Pomeroy, K. (2014) Pomeroy's ratings. Available: http://www.kenpom.com
[14]. Sagarin, J. (2014). Basketball Statistics. Available: http://www.usatoday30.usatoday.com/sports/sagarin.htm.
[15]. Schwertman, Neil C.; McCready,Thomas A.; and Howard, Lesley (1991). *Probability Models for the NCAA Regional Basketball Tournaments.* The American Statistician, 45(1): 35-38.
[16]. Shen, Gang; Hua, Su;Zhang, Xiao; Mu, Yingfei;and Magel, Rhonda (2015). *Predicting Results of March Madness Using the Probability Self-Consistent Method.* International Journal of Sports Science, 2015, 5(4): 139-144.
[17]. Wang, Wenting, and Magel, Rhonda (2014). *Predicting Winners of NCAA Women's Basketball Tournament Games*. International Journal of Sports Science, 2014, 4(5): 173-180.
[18]. West, B.T. (2006). *A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament.* Journal of Quantitative Analysis in Sports, 2, 3, p 3-8.
[19]. West, B.T. (2008). *A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament: Updated Results from 2007.* Journal of Quantitative Analysis in Sports, 4, 2, p 6-8.